Comparing performance of biometric models between different groups with Bayesian statistics

Alen Ajanović
Peter Peer
Žiga Emeršič

Fakulteta za računalništvo in informatiko
Univerza v Ljubljani

Introduction

- How general are the models we build?
  - Voice recognition?
  - Face detection?

- Many instances where a model doesn't work on a different group than it was trained on
- Is the same true for ears?
  - What about neural networks?

Data

- 2018/19 ear dataset we have built during this course
  - *It is not without its flaws*
- We have used a pre-trained haar cascade model
- And three separate neural network models
  - Trained on females (1.977 images)
  - Trained on males (5.984 images)
  - Trained 70/30 split (10.214 images)

Methodology

- Each model made predictions for random images (both males/females)
- Both groups described by a 250-length IoU vector
  - Each IoU reading measured with 200 random images

#reading	1	2	3		250
IoU	0.42	0.32	0.35	...	0.74

Male IoU vector

#reading	1	2	3		250
IoU	0.45	0.29	0.46	...	0.63

Female IoU vector

Methodology (Bayes)

#reading	1	2	3		250
IoU	0.42	0.32	0.35	...	0.74

= y

- In Bayesian statistics we describe our prior beliefs

$$\mu = N(70, 20)$$

$$\sigma = U(0, 1)$$

$$y|(\mu, \sigma) = N(\mu, \sigma)$$

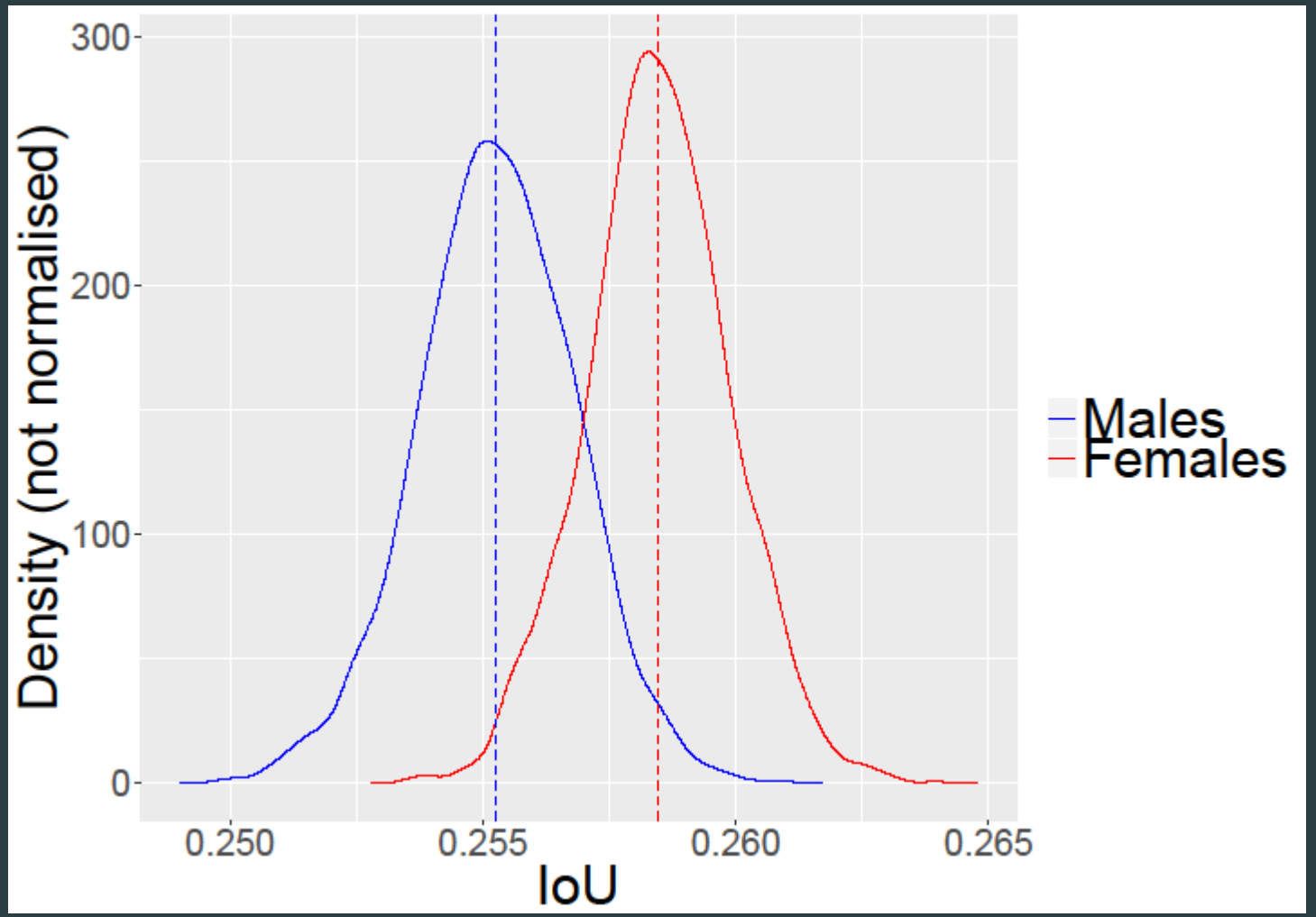
- Result: $\mu_{posterior}|y$, $\sigma_{posterior}|y$

Methodology (Bayes)

- Result: $\mu_{posterior}$, $\sigma_{posterior}$

- In fact, we obtain many possible values, not just one (we sample from the posterior distribution)
- In our case we obtain 4.000 samples of both parameters, but really only care about the mean
- Perhaps better illustrated on results

Results (haar)

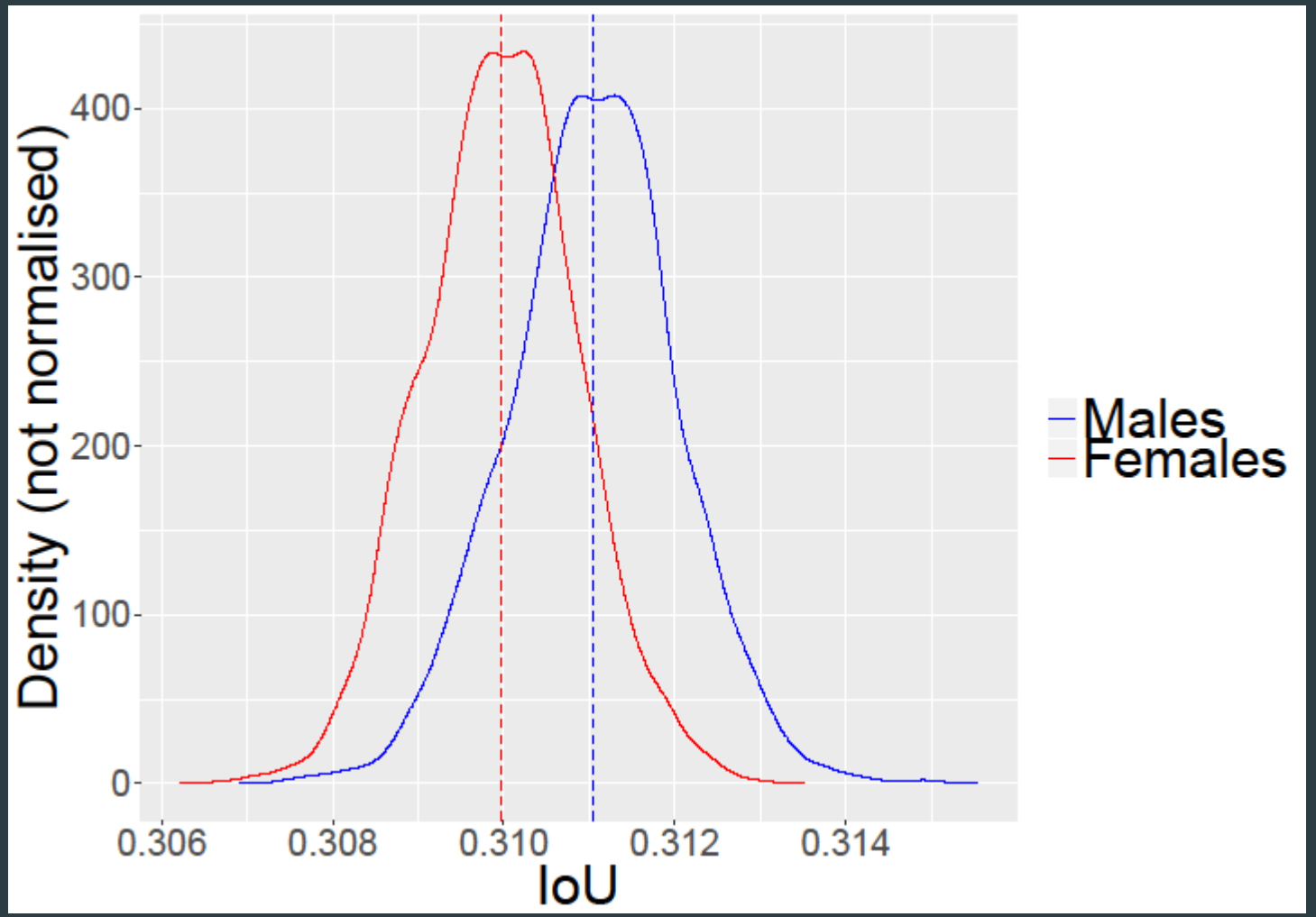


Results (total IoU)

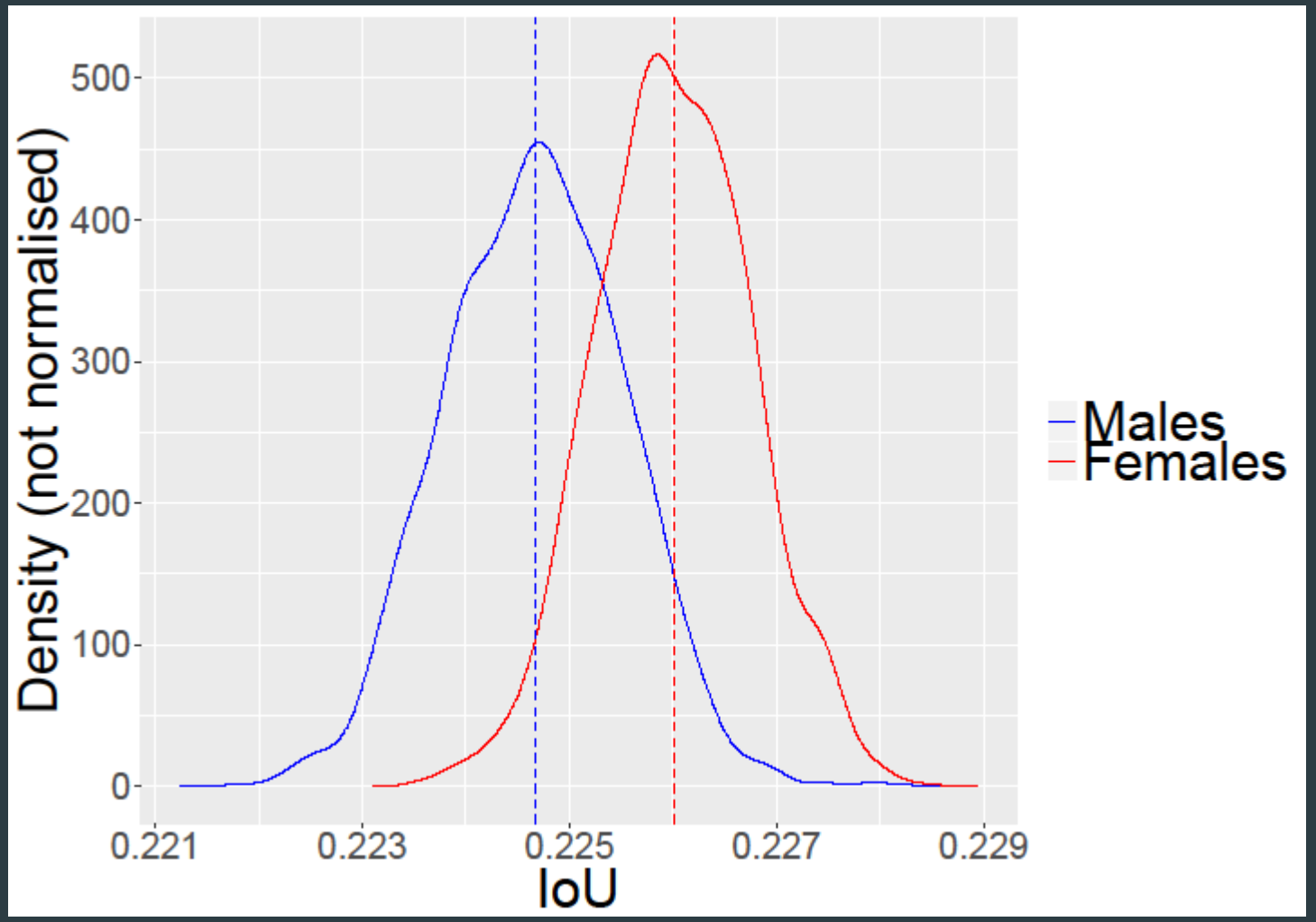
Haar	70/30 NN	Female NN	Male NN
0,255	0,338	0,236	0,333

- Better result is correlated with a bigger training set
- No major difference between only training on males vs. males and females

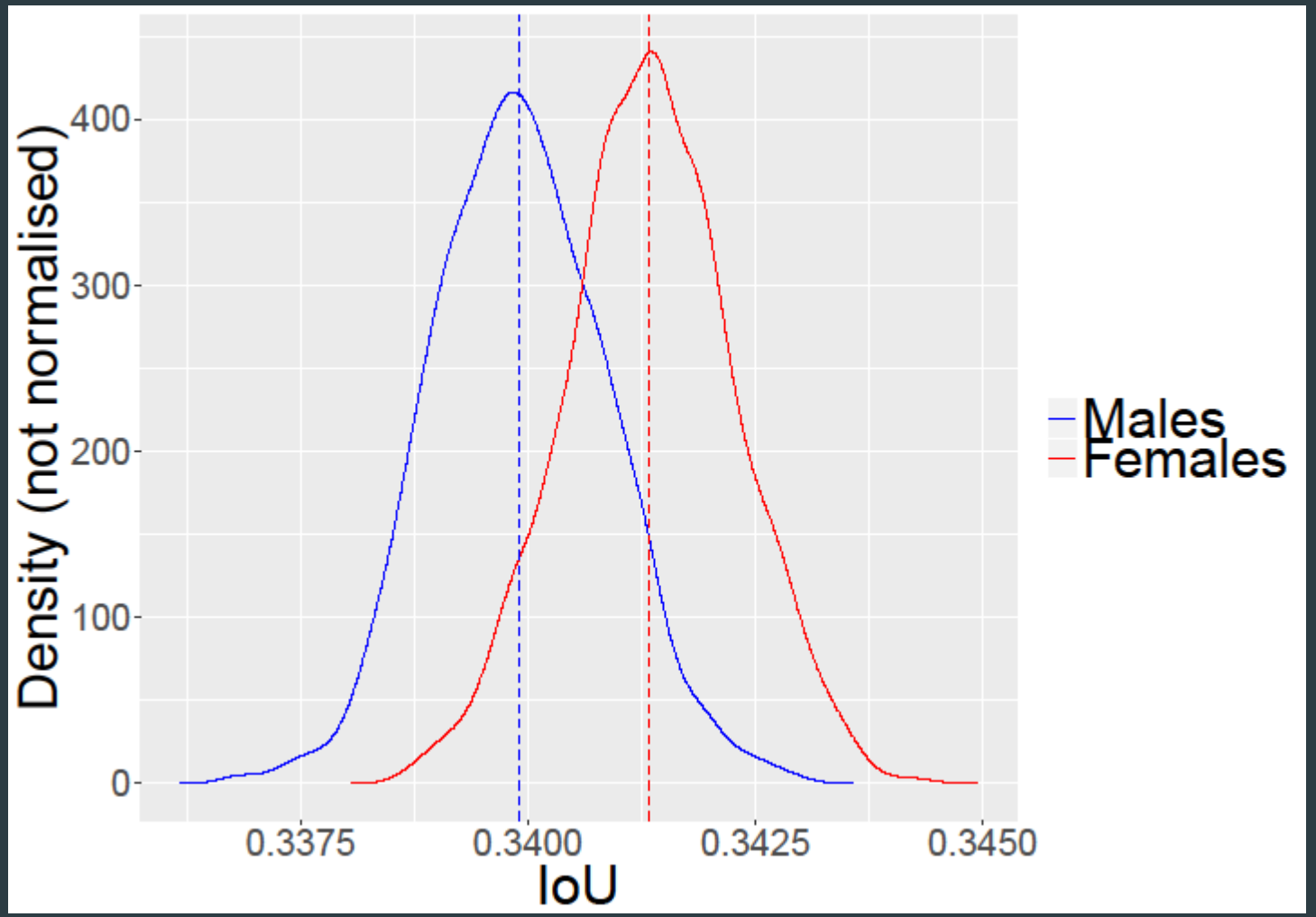
Results (70/30 NN)



Results (Female NN)



Results (Male NN)



Conclusion

- Bias exists, but is practically insignificant (<1%)
- The models seem to perform better on women, regardless of the initial training set (maybe biological reasons?)
- Training a network only on males does not inhibit its performance on females if training set is large enough
- Even still, the areas of non-intersection seem large enough to be statistically significant, meaning bias must come from somewhere (perhaps just variance introduced in our methodological process)

Questions?